JOURNAL OF SIMULATION
https://doi.org/10.1080/17477778.2021.1955629

Efficient confidence intervals for the difference of two Bernoulli distributions' success parameters

Ignacio Erazo and David Goldsman

Georgia Institute of Technology, Atlanta, GA, U.S.A

ABSTRACT

We study properties of confidence intervals (CIs) for the difference of two Bernoulli distributions' success parameters. The CIs under investigation range from the classical fixed-sample-size CI to sequential versions, possibly incorporating batching. For each CI method, we examine the attained coverage, as well as the trade-offs between the number of observations and stages required to obtain a desired CI width. We consider cases in which the two populations are completely independent, and we provide analytical and simulation results to measure the performance of the different methods. For the multi-stage methods, we find that a simple observation allocation rule based on comparing the sample standard deviations of the two populations is more efficient than taking equal sample sizes from both. We also show that the use of a moderate level of batching saves stages at only modest costs in sample size and coverage.

ARTICLE HISTORY
Received 10 March 2021
Accepted 07 July 2021

KEYWORDS
Confidence intervals; Bernoulli success parameters; two-sample differences; simulation

1. Introduction and motivation

Our interest in this paper lies in obtaining parsimonious sequential confidence intervals (CIs) for the difference of the success parameters arising from two Bernoulli distributions. This goal can be motivated by several practical applications.

- Given a highly dangerous and virulent pandemic that is circulating around the world, a university wants to compare the proportion of students currently infected with the disease to the proportion of faculty infected.
- A pharmaceutical company is interested in studying the efficacy of a new drug, and so will compare the probability that the new drug provides immunity vs. that of a placebo.
- A warehouse logistics manager wants to know which of two simulated inventory policies has the higher probability of yielding a 98% on-time delivery rate of goods to a client, where a "success" is defined as the event that at least 98% of all orders were delivered on time during a particular simulation replication.

Of course, sampling can be expensive, so in each of the above scenarios, it is important to allocate the available observational units wisely. How can this be accomplished?

To put things on a more-solid footing, suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid) Bern(p_x) random variables, $Y_1, Y_2, \ldots, Y_m$ are iid Bern(p_y), and that the X's and Y's are independent. Such iid observations are easily obtained in the context of computer simulation by running independent replications of the each of two simulation models.

We first consider the "classical" Wald approximate CI for $p_x - p_y$ of the form

$$p_x - p_y \in \bar{X} - \bar{Y} \pm H \equiv \bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n} + \frac{\bar{Y}(1-\bar{Y})}{m}}, \quad (1)$$

where $\bar{X} \equiv \sum_{i=1}^{n} X_i/n$ and $\bar{Y} \equiv \sum_{j=1}^{m} Y_j/m$ are the respective sample means, $z_{\alpha/2}$ denotes the usual $(1 - \frac{\alpha}{2})$ quantile of the standard normal distribution, and the half-length H is implicitly defined. It is well-known that this CI has poor small-sample properties. In particular, for small values of n and m, the coverage probability of this CI can substantially undershoot or even overshoot the nominal probability $1 - \alpha$ (see, e.g., the surveys in Agresti and Coull (1998); Brown et al. (2001, 2002); Frey (2010), and Newcombe (1998) for the analogous single-parameter Bern(p) case). Various techniques exist to push the coverage towards the nominal value. For instance, for the single-parameter case, Agresti and Coull (1998) and Wilson (1927) study fixed-sample-size CIs that append correction terms to Wald's CI, and Clopper and Pearson (1934) give an exact CI obtained by inverting certain binomial tests.

One often wishes to produce a CI of at most a certain length. This task typically requires more than one stage of observations to be carried out. Such procedures in the context of a single

CONTACT David Goldsman sman@gatech.edu H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0205, U.S.A

Bernoulli parameter are studied by, e.g., Armitage (1958), Khan (1998), Robbins and Siegmund (1974), Tanaka (1961), and Yaacoub et al. (2019b) (which gives a certain optimal sampling rule), and Zacks and Mukhopadhyay (2007). There are other criteria that dictate when to stop sampling besides delivering a particular half-width. For instance, one can consider the *proportional accuracy* criterion, where the CI for p in the one-sample case is given by $\{p : |\bar{X} - p| < cp\}$ for a fixed $c \in (0, 1)$ (Huber (2017) and Malinovsky and Zacks (2018)). See Mukhopadhyay and Banerjee (2015) for yet a different intuitive stopping criterion. Turner et al. (2013) illustrates the estimation of the Bernoulli success parameter (and others) in the context of simulation. In terms of transitioning the single-sample Bern(p) CIs over to a two-sample version, the current paper begins with the basic CI (1) (with none of the corrections mentioned above) and proposes two-stage and sequential heuristics in order to satisfy a certain bound on the CI half-width. One could also invoke a ranking-and-selection perspective on the problem of finding that one of $k \geq 2$ Bernoulli populations having the largest (or smallest) success parameter (see, e.g., Bechhofer et al. (1995) and Goldsman (2015)); but this approach is not considered here.

This paper studies analytical and simulation-based methodologies to determine how many and what additional observations to take, starting from something like (1), in order to obtain a CI of at most a pre-specified length. The organisation of the article is as follows. §2 discusses exact performance properties of the basic CI (1). In §3, we analyse an "optimized" two-stage CI procedure that attempts to efficiently allocate the available budget so as to produce a CI that approximately satisfies a user-specified half-width requirement; the procedure is "naive" in the sense that it does not *guarantee* the half-width requirement. §4 is concerned with sequential sampling rules. In particular, §4.1 considers a one-at-a-time sequential sampling procedure, while §4.2 proposes a compromise in which we *batch* observations – thus potentially adding a few extra observations, but certainly reducing the number of sampling stages (which corresponds to a time savings). §5 presents a Monte Carlo evaluation of the various procedures, where we report on performance measures such as CI coverage probability and expected half-length, as well as the expected number of stages (and observations) required to obtain the pre-specified half-width. We find that batching provides a good middle ground that offers savings in terms of both observations and stages. §6 gives a summary of our findings as well as suggestions for future study.

2. Properties of the basic confidence interval (1)

We first consider performance characteristics for the easiest case – the basic single-stage confidence interval for $p_x - p_y$ of the form (1). Its coverage probability and expected half-width can be computed exactly for any given p_x, p_y, n, m, and α.

We continue to assume that the X's and Y's are independent samples, and we define

$$\begin{aligned} g(i,j) &\equiv \Pr(\bar{X} = i/n, \bar{Y} = j/m) \\ &= \Pr(\bar{X} = i/n) \Pr(\bar{Y} = j/m) \\ &= \binom{n}{i} p_x^i q_x^{n-i} \binom{m}{j} p_y^j q_y^{m-j}, \\ &\quad i = 0, 1, \ldots, n, j = 0, 1, \ldots, m, \end{aligned}$$

where the second equality comes from independence and the third from the binomial distribution, and where $q_x \equiv 1 - p_x$ and $q_y \equiv 1 - p_y$. On the way to an expression for the exact coverage, we denote the indicator function for an arbitrary event $\mathcal{E}$ by

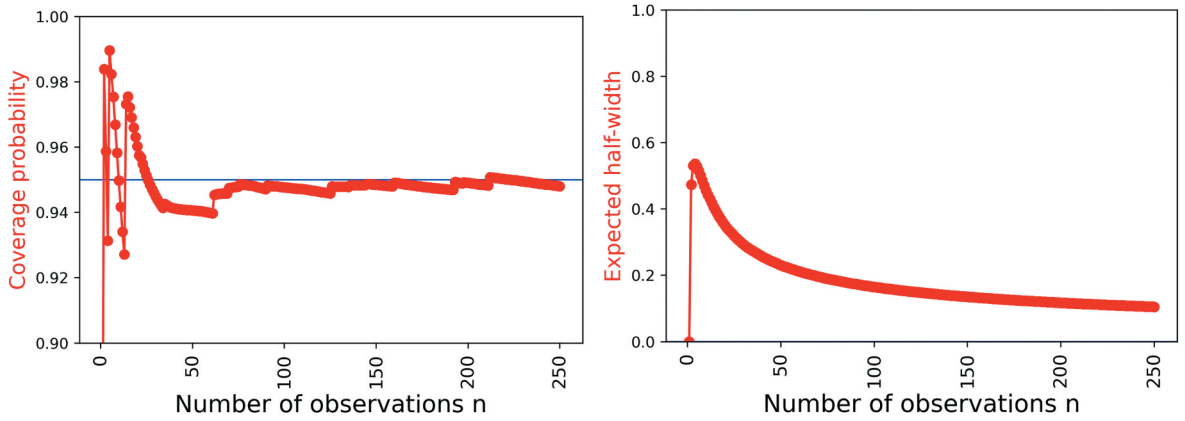
$$1(\mathcal{E}) \equiv \begin{cases} 1 & \text{if } \mathcal{E} \text{ occurs} \\ 0 & \text{otherwise.} \end{cases}$$

In addition, we define the event $\mathcal{C} \equiv \{\text{CI covers } p_x - p_y\}$. Then exact expressions for the coverage and expected half-width of the CI are given by

$$\begin{aligned} \Pr(\mathcal{C}) &= \mathrm{E}[1(\mathcal{C})] \\ &= \sum_{i=0}^{n} \sum_{j=0}^{m} 1\left(\mathcal{C} \,|\bar{X} = \frac{i}{n}, \bar{Y} = \frac{j}{m}\right) g(i,j) \\ &= \sum_{i=0}^{n} \sum_{j=0}^{m} 1\left(p_x - p_y \in \bar{X} - \bar{Y} \pm H | \bar{X} = \frac{i}{n}, \bar{Y} = \frac{j}{m}\right) g(i,j) \\ &= \sum_{i=0}^{n} \sum_{j=0}^{m} 1\left(p_x - p_y \in \frac{i}{n} - \frac{j}{m} \pm z_{\alpha/2} \sqrt{\frac{i}{n^2}\left(1 - \frac{i}{n}\right) + \frac{j}{m^2}\left(1 - \frac{j}{m}\right)}\right) g(i,j), \end{aligned} \tag{2}$$

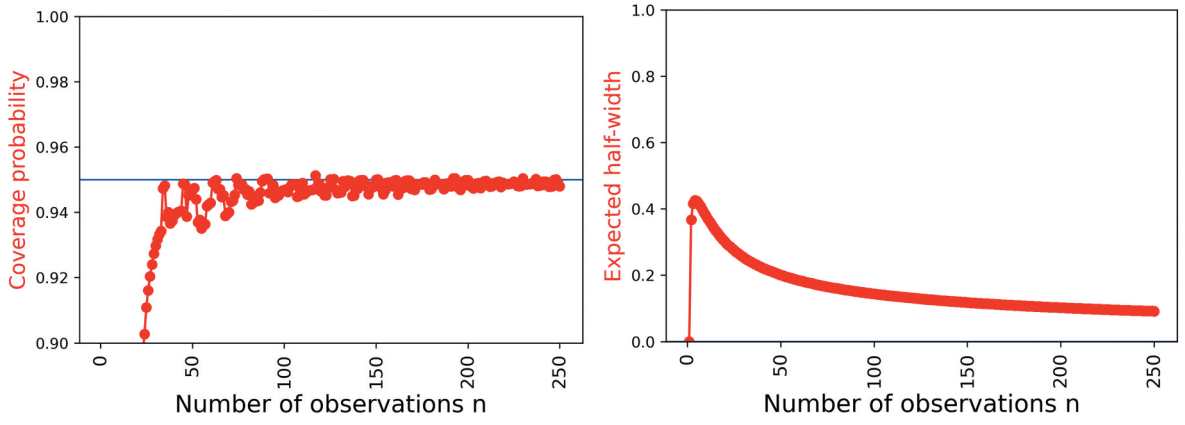
and

$$\begin{aligned} \mathrm{E}[H] &= \sum_{i=0}^{n} \sum_{j=0}^{m} \left[H \,|\bar{X} = \frac{i}{n}, \bar{Y} = \frac{j}{m}\right] g(i,j) \\ &= z_{\alpha/2} \sum_{i=0}^{n} \sum_{j=0}^{m} \sqrt{\frac{i}{n^2}\left(1 - \frac{i}{n}\right) + \frac{j}{m^2}\left(1 - \frac{j}{m}\right)} g(i,j). \end{aligned} \tag{3}$$



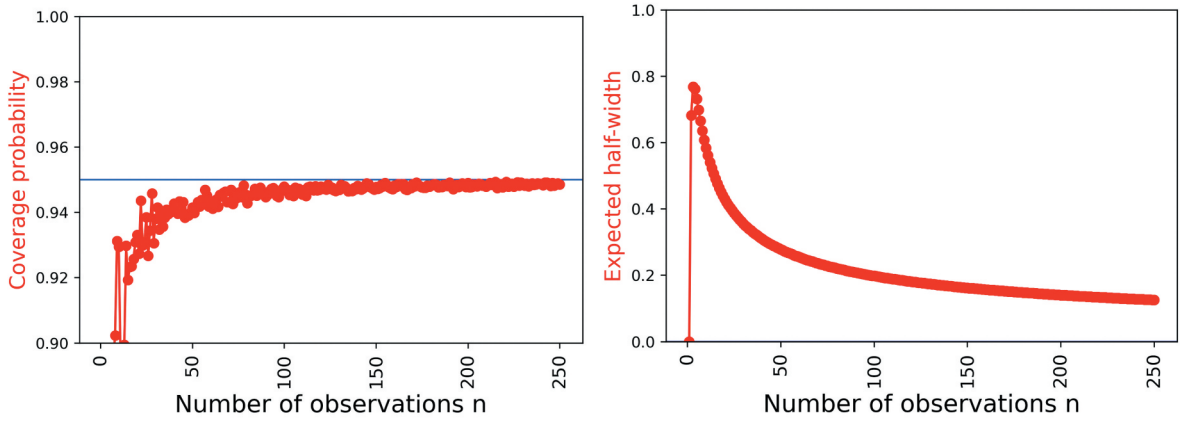
(a) Coverage for $p_x = p_y = 0.1$

(b) $\mathrm{E}[H]$ for $p_x = p_y = 0.1$



(c) Coverage for $p_x = 0.1,\ p_y = 0.05$

(d) $\mathrm{E}[H]$ for $p_x = 0.1,\ p_y = 0.05$



(e) Coverage for $p_x = 0.3,\ p_y = 0.05$

(f) $\mathrm{E}[H]$ for $p_x = 0.3,\ p_y = 0.05$

Figure 1. Coverage and expected half-width of the basic confidence interval (1), for three choices of probability pairs (p_x, p_y), confidence level $100(1-\alpha) = 95\%$, and equal sample sizes $n = m = 1, 2, \ldots, 250$.

Figure 1 depicts exact results based on Equations (2) and (3) for, respectively, the coverage probability and expected half-width of the basic confidence interval (1) for three choices of probability pairs (p_x, p_y), confidence level $100(1-\alpha) = 95\%$, and all equal sample sizes $n = m$ up to 250. We see that the expected half-width exhibits reasonably smooth behaviour; and when the sample size is above 5, it is monotonically decreasing. The achieved coverage is quite a bit more erratic; most of the time it falls

below the desired value of 95% for small n (which is a known issue for (1)). Notably, when $p_x \approx p_y$, the coverage exhibits local peaks with respect to n, followed by intervals where the coverage temporarily decreases as the sample size increases. Of course, as n becomes large, the coverage approaches the nominal value $1-\alpha$.

3. Two-stage confidence interval

This section discusses a simple two-stage procedure based on the classical CI (1). As is typical of such procedures, the first stage of sampling establishes a baseline CI for $p_x - p_y$, which is subsequently used to suggest a second-stage sample size that will yield a CI that (approximately) has a desired half-width. The procedure is "naive" in the sense that the delivered half-width is random and does not necessarily meet the specification.

In order to produce preliminary estimates of p_x and p_y, suppose that we have taken an initial (first-stage) sample of n_1 iid Bern(p_x) X's and m_1 iid Bern(p_y) Y's, where the X's are independent of the Y's, and n_1 and m_1 are specified beforehand by the user. These observations could be obtained, for instance, from n_1 and m_1 independent replications of two simulated processes. Let $c_x \equiv \bar{X}_1(1-\bar{X}_1)$ and $c_y \equiv \bar{Y}_1(1-\bar{Y}_1)$, where $\bar{X}_1$ and $\bar{Y}_1$ are the respective sample means from the initial stage of n_1 and m_1 observations. Then, the initial CI based on our first-stage samples of sizes n_1 and m_1 is

$$p_x - p_y \in \bar{X}_1 - \bar{Y}_1 \pm z_{\alpha/2}\sqrt{\frac{c_x}{n_1} + \frac{c_y}{m_1}}.$$

3.1. Optimised sample-size calculation

With our limited knowledge about the state of affairs after just one stage of observations, we would like to decide how many extra observations we need to take in order to reduce our confidence interval's half-width H to a value that is at most some specified error ϵ. Specifically, if n and m are the total sample sizes of the X's and Y's, respectively, what is the smallest overall budget $N = n + m$ that will "likely" give us a CI having half-width $H \le \epsilon$?

To simplify things a little bit in the upcoming discussion, let $n = \beta N$ and $m = (1-\beta)N$ for some suitable $0 < \beta < 1$, where we temporarily ignore the possibility that n and m might not be integers. Then, the goal is to find some "optimal" N and β, so that

$$z_{\alpha/2}\sqrt{\frac{c_x}{\beta N} + \frac{c_y}{(1-\beta)N}} \le \epsilon,$$

or, equivalently,

$$N \ge \frac{z_{\alpha/2}^2}{\epsilon^2}\left[\frac{c_x}{\beta} + \frac{c_y}{1-\beta}\right]. \tag{4}$$

We will minimise this quantity with respect to β. To this end, set

$$f(\beta) \equiv \frac{c_x}{\beta} + \frac{c_y}{1-\beta}$$

and take

$$f'(\beta) = \frac{-c_x}{\beta^2} + \frac{c_y}{(1-\beta)^2} = 0.$$

Notice that if both c_x and c_y are equal to zero, then $f(\beta) = 0$. So we henceforth assume that at least one of them is greater than zero.

(I) If $c_x = 0$ and $c_y > 0$ [$c_y = 0$ and $c_x > 0$], then since $\beta \in (0,1)$, we have $f'(\beta) > 0$ [$f'(\beta) < 0$], and the minimum is achieved when β approaches zero [one].

(II) When both c_x and c_y are greater than zero, we solve $f'(\beta) = 0$, which means solving $\beta^2(c_y - c_x) + 2\beta c_x - c_x = 0$. When $c_x = c_y$, it is clear that the solution is $\beta = 0.5$. Otherwise, there are two solutions, but the only critical point that is between 0 and 1 is

$$\beta^\star \equiv \frac{-c_x + \sqrt{c_x c_y}}{c_y - c_x} = \frac{\sqrt{c_x}}{\sqrt{c_x} + \sqrt{c_y}} = \frac{1}{1 + \sqrt{c_y/c_x}},$$

which is determined by the ratio of the (non-zero) estimated standard deviations of the X's and Y's.

We also see that the second derivative of $f(\beta)$ is

$$f''(\beta) = \frac{2c_x}{\beta^3} + \frac{2c_y}{(1-\beta)^3}.$$

This quantity is positive for $\beta \in (0,1)$, so $\beta^\star$ yields the minimum of $f(\beta)$ for all cases. Then, Equation (4) suggests that we will need N to be at least

$$N \ge N^\star \equiv \frac{z_{\alpha/2}^2}{\epsilon^2}\left[\frac{c_x}{\beta^\star} + \frac{c_y}{1-\beta^\star}\right] = \frac{z_{\alpha/2}^2\left(\sqrt{c_x} + \sqrt{c_y}\right)^2}{\epsilon^2}. \tag{5}$$

The quantity $\beta^\star$ is used to divide the number of observations N between the two populations, so that the total sample sizes will be $n = \langle \beta^\star N^\star \rangle$ and $m = \langle (1-\beta^\star)N^\star \rangle$, where $\langle \cdot \rangle$ denotes the "round-to-the-nearest-integer" function. Note that c_x and c_y are bounded from above by $1/4$, resulting in a conservative upper bound for the total sample size of $N^\star \le z_{\alpha/2}^2/\epsilon^2$.

3.2. Exact analysis

The optimised two-stage approach comes with at least two potential drawbacks: Since c_x and c_y are random variables (before observations are taken), we see that (i) the second-stage sample size is a random variable, and (ii)

there is no guarantee that the $N^\star$ and $\beta^\star$ described above will actually result in a CI having a half-width $\leq \epsilon$.

We will illustrate these issues by carrying out some exact calculations under the assumption that the first-stage sample sizes are n_1 and m_1. From Equation (5), we immediately have that the naive two-stage heuristic to obtain a confidence interval having half-width $H \leq \epsilon$ suggests a total of

$$N \equiv \left\lceil \frac{z^2_{\alpha/2}}{2} \left[\sqrt{\bar{X}_1(1-\bar{X}_1)} + \sqrt{\bar{Y}_1(1-\bar{Y}_1)} \right]^2 \right\rceil$$

observations, where "$\lceil \cdot \rceil$" denotes the "ceiling" (integer round-up) function. In fact, in light of the first-stage observations, the total number of observations taken is (a slightly redefined) $N^\star \equiv \max\{n_1+m_1, N\}$.

We define the related quantity

$$N(i,j) \equiv \left\lceil \frac{z^2_{\alpha/2}}{\epsilon^2} \left[\sqrt{\frac{i}{n_1}\left(1-\frac{i}{n_1}\right)} + \sqrt{\frac{j}{m_1}\left(1-\frac{j}{m_1}\right)} \right]^2 \right\rceil$$

which is valid for $i = 0, 1, \ldots, n_1$ and $j = 0, 1, \ldots, m_1$; so $N(i,j)$ is just N given that $\bar{X}_1 = i/n_1$ and $\bar{Y}_1 = j/m_1$. We also define $n^\star(i,j) \equiv \langle \max\{n_1, N(i,j)\beta^\star\}\rangle$ and $m^\star(i,j) \equiv \langle \max\{m_1, N(i,j)(1-\beta^\star)\}\rangle$ for all i,j, where $\beta^\star$ is chosen as in §3.1. Both quantities represent the total numbers of observations we end up taking for each population given the first-stage results. We also define the corresponding second-stage sample sizes, $n_2(i,j) \equiv n^\star(i,j) - n_1$ and $m_2(i,j) \equiv m^\star(i,j) - m_1$ for all i,j.

Similar to the notation introduced in §2, we define the first-stage sample probabilities for all $i = 0, 1, \ldots, n_1$ and $j = 0, 1, \ldots, m_1$,

$$\begin{aligned} g_1(i,j) &= \Pr(\bar{X}_1 = i/n_1, \bar{Y}_1 = j/m_1) \\ &= \binom{n_1}{i} p_x^i q_x^{n_1-i} \binom{m_1}{j} p_y^j q_y^{m_1-j}. \end{aligned}$$

The numbers of successes for the two respective populations in the second stage are given by

$$T_x(i,j) \equiv \sum_{r=n_1+1}^{n^\star(i,j)} X_r \quad \text{and} \quad T_y(i,j) \equiv \sum_{r=m_1+1}^{m^\star(i,j)} Y_r;$$

and we define the second-stage sample probabilities for all $i = 0, 1, \ldots, n_1$, $j = 0, 1, \ldots, m_1$, $k = 0, 1, \ldots, n_2(i,j)$, and $\ell = 0, 1, \ldots, m_2(i,j)$ by

$$\begin{aligned} g_2(i,j,k,\ell) &= \Pr(T_x(i,j) = k, T_y(i,j) = \ell) \\ &= \binom{n_2(i,j)}{k} p_x^k q_x^{n_2(i,j)-k} \binom{m_2(i,j)}{\ell} p_y^\ell q_y^{m_2(i,j)-\ell}. \end{aligned}$$

With all of this notation in mind, it is easy to see that the expected total number of observations to be taken from the two populations in the naive two-stage procedure is

$$\mathrm{E}[N^\star] = \sum_{i=0}^{n_1} \sum_{j=0}^{m_1} (n^\star(i,j) + m^\star(i,j)) g_1(i,j).$$

We now obtain expressions for the two-stage procedure's coverage probability and expected half-width. Recall that the event $\mathcal{C} = \{\text{CI covers } p_x - p_y\}$, so that

$$\begin{aligned} \Pr(\mathcal{C}) &= \mathrm{E}[1(\mathcal{C})] \\ &= \sum_{i=0}^{n_1} \sum_{j=0}^{m_1} 1\left(\mathcal{C} \,|\, \bar{X}_1 = \frac{i}{n_1}, \bar{Y}_1 = \frac{j}{m_1}\right) g_1(i,j) \\ &= \sum_{i=0}^{n_1} \sum_{j=0}^{m_1} \sum_{k=0}^{n_2(i,j)} \sum_{\ell=0}^{m_2(i,j)} 1\Bigg(\mathcal{C} | \bar{X}_1 = \frac{i}{n_1}, \bar{Y}_1 = \frac{j}{m_1}, \\ &\qquad T_x(i,j) = k, T_y(i,j) = \ell\Bigg) g_2(i,j,k,\ell) g_1(i,j) \\ &= \sum_{i=0}^{n_1} \sum_{j=0}^{m_1} \sum_{k=0}^{n_2(i,j)} \sum_{\ell=0}^{m_2(i,j)} 1\Bigg(p_x - p_y \in \bar{X} - \bar{Y} \pm H_{ijk\ell} | \\ &\qquad \bar{X} = \frac{i+k}{n^\star(i,j)}, \bar{Y} = \frac{j+\ell}{m^\star(i,j)}\Bigg) g_2(i,j,k,\ell) g_1(i,j), \end{aligned}$$

where $\bar{X}$ and $\bar{Y}$ are the total sample means taken over both stages, and

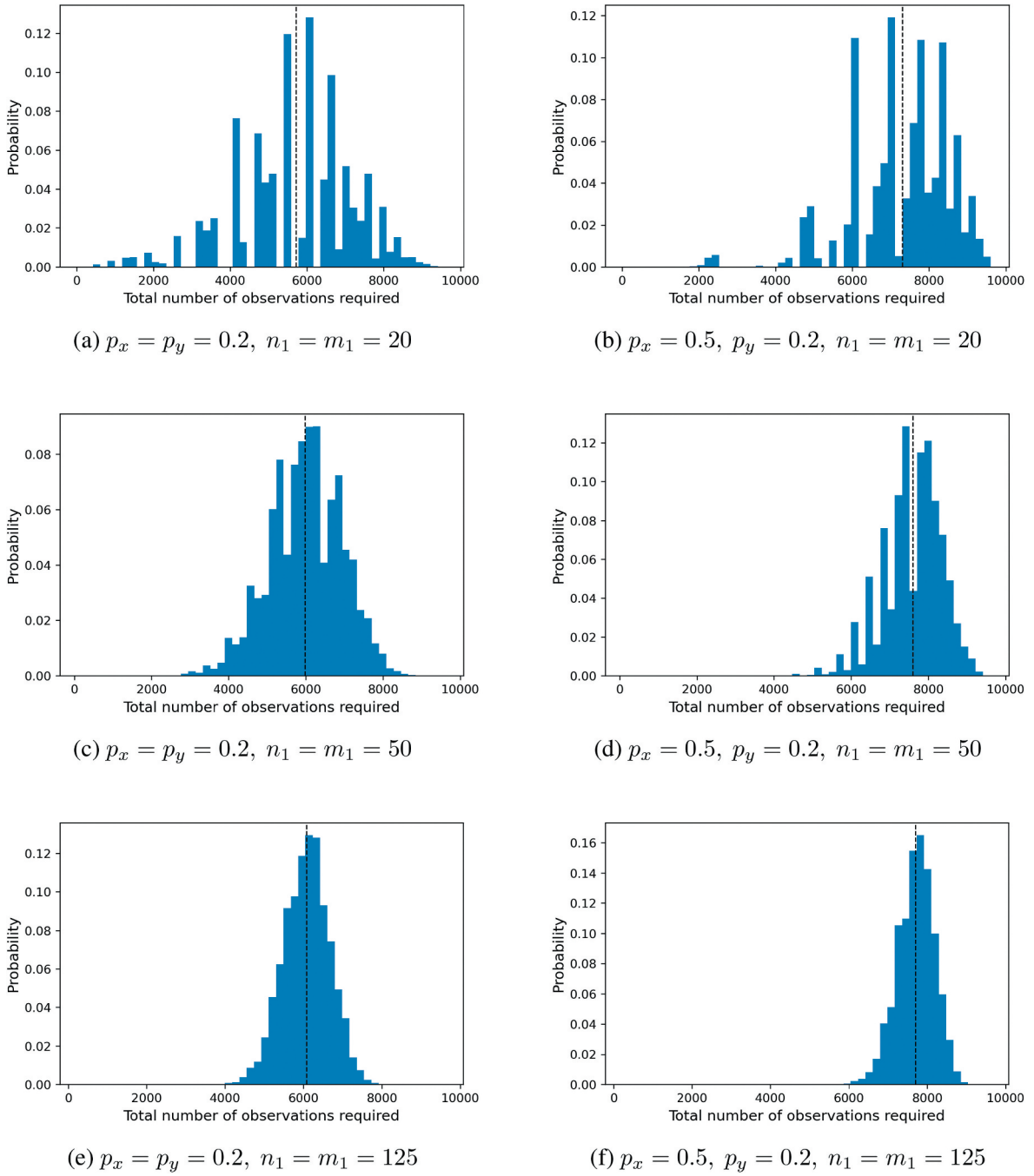
$$H_{ijk\ell} \equiv z_{\alpha/2} \sqrt{\frac{i+k}{n^\star(i,j)^2}\left(1-\frac{i+k}{n^\star(i,j)}\right) + \frac{j+\ell}{m^\star(i,j)^2}\left(1-\frac{j+\ell}{m^\star(i,j)}\right)}.$$

Similarly, we can also calculate the expected half-width of the delivered CI,

$$\mathrm{E}[H] = \sum_{i=0}^{n_1} \sum_{j=0}^{m_1} \sum_{k=0}^{n_2(i,j)} \sum_{\ell=0}^{m_2(i,j)} H_{ijk\ell}\, g_2(i,j,k,\ell) g_1(i,j).$$

These equations give us straightforward ways to calculate the expected sample size, coverage probability, and expected half-width for fixed n_1, m_1, p_x, p_y, and α. If the desired error ϵ is very small and/or the probabilities are near 0.5, then more "work" is required to obtain the confidence interval. This extra work might be in the form of prohibitively large sample sizes used in the calculations, in which case simulation may be necessary to estimate the three performance measures.

Figure 2 depicts histograms of the number of observations $N^\star$ required by our naive two-stage procedure for probability pairs $(p_x, p_y) = (0.2, 0.2)$ and $(0.5, 0.2)$; equal first-stage sample sizes $n_1 = m_1 = 20$ (small-sample case), 50 (medium), and 125 (large); desired half-width $\epsilon = 0.02$; and nominal confidence level 95%. It is clear that $N^\star$'s variability is significant for small initial sample sizes. In addition, more observations are required when either p_x or $p_y \approx 0.5$.



(a) $p_x = p_y = 0.2,\ n_1 = m_1 = 20$ (b) $p_x = 0.5,\ p_y = 0.2,\ n_1 = m_1 = 20$

(c) $p_x = p_y = 0.2,\ n_1 = m_1 = 50$ (d) $p_x = 0.5,\ p_y = 0.2,\ n_1 = m_1 = 50$

(e) $p_x = p_y = 0.2,\ n_1 = m_1 = 125$ (f) $p_x = 0.5,\ p_y = 0.2,\ n_1 = m_1 = 125$

Figure 2. Total number of observations required by the two-stage procedure for $(p_x, p_y) = (0.2, 0.2)$ and $(0.5, 0.2)$; $n_1 = m_1 = 20$, 50, and 125; $\epsilon = 0.02$; and $1 - \alpha = 0.95$. The vertical line represents $\mathsf{E}[N^\star]$.

4. Sequential sampling procedures

The naive two-stage procedure may suffer from a number of potential drawbacks. For instance,

•The procedure's required sample size can be quite random, and is strongly dependent on the first-stage sample sizes (n_1, m_1) (Figure 2).

• The observations are not used efficiently (since there is only one chance at the end of the first stage to make a sampling decision).

• It does *not guarantee* the delivery of a half-width $\leq \epsilon$; and, in addition, the delivered half-width H has high variability.

• The coverage probability achieved is not always $1 - \alpha$, and with smaller sample sizes tends to be below that nominal value; see the discussion in §5.

This section aims to take advantage of sequential sampling to mitigate the above issues. We first present in §4.1 a fully sequential one-at-a-time sampling heuristic, followed by a procedure in §4.2 that incorporates *batching* to save on sampling stages – which is useful since stages can be interpreted as "time".

4.1. Fully sequential procedure

In this subsection, we consider a one-at-a-time sequential sampling procedure. The idea is that after having taken a preliminary sample (n_1, m_1), we will execute a sequential procedure to eventually obtain a CI of half-width $\leq \epsilon$. We will take observations one-at-a-time in a "greedy" way that aims to get the biggest bang for the buck by deciding to take an observation from X or Y based on which is "more likely" to reduce the CI width the most. We stop sampling and deliver our CI after the first stage for which the half-width is $\leq \epsilon$.

We denote the sample means after r stages of observations have been taken by $\bar{X}_r$ and $\bar{Y}_r$, along with respective cumulative sample sizes of n_r and m_r, $r = 1, 2, \ldots$. We propose a simple sampling rule based roughly on a comparison of the sample standard errors of the X's and Y's. Namely, while the CI still has half-width $> \epsilon$, continue sampling in such a way as to attempt to reduce the half-width as much as possible: If at the end of stage $r-1$ we have

$$\begin{aligned} &\bar{X}_{r-1}(1-\bar{X}_{r-1})\left[\frac{1}{n_{r-1}} - \frac{1}{n_{r-1}+1}\right] \\ &> \bar{Y}_{r-1}(1-\bar{Y}_{r-1})\left[\frac{1}{m_{r-1}} - \frac{1}{m_{r-1}+1}\right], \end{aligned} \tag{6}$$

then take the next observation (stage r) from the X's because we are likely to get more bang for the buck by that choice; otherwise, take from the Y's. In the case of equality, as both samples presumably have the same cost, the tie is broken arbitrarily. Continue until we finally meet the desired half-width criterion. The greedy optimisation heuristic given by (6) makes sense because, at any stage, the left- and right-hand sides are rough measures of the incremental CI width reductions to be obtained by taking the next observation from X or Y, respectively.

We initialise the first-stage sample sizes for the two populations by n_1 and m_1. If, on any given subsequent stage $r = 2, 3, \ldots$, we decide to take an X observation, then we set $n_r \leftarrow n_{r-1} + 1$ and $m_r \leftarrow m_{r-1}$; else take a Y observation and set $n_r \leftarrow n_{r-1}$ and $m_r \leftarrow m_{r-1} + 1$.

In the next section, simulation will be used to approximate the coverage probability, the expected CI half-width, and the expected number of observations needed to obtain a CI of the desired width.

4.2. Sequential procedure with batching

In practice, the strategy of executing observations one-at-a-time, and then waiting for the result of that stage in order to proceed with the next one, is often impractical because of time expenditures. Instead, we can consider the middle-ground scenario in which *batches* of B observations-at-a-time are taken and processed before the next batch (stage) starts. This is reminiscent of a testing protocol for epidemics, where B test observations can be performed each day, with results typically available after 24–48 hours. Of course, the fully sequential procedure described in §4.1 is simply the $B = 1$ special case.

We partially repeat previous notation definitions, this time adapted for batching. We again initialise the first-stage sample sizes for the two populations as n_1 and m_1. Suppose that, on any given subsequent stage $r = 2, 3, \ldots$, we decide to take a batch of B observations, divided into b_r X observations and $B - b_r$ Y's. (We describe how to choose b_r below.) Then, we set $n_r \leftarrow n_{r-1} + b_r$ and $m_r \leftarrow m_{r-1} + B - b_r$ and let

$$\bar{X}_r \equiv \frac{1}{n_r}\sum_{i=1}^{n_r} X_i \quad \text{and} \quad \bar{Y}_r \equiv \frac{1}{m_r}\sum_{i=1}^{m_r} Y_i$$

denote the cumulative sample means of the X's and Y's after r stages of sampling have been completed. Moreover, we also denote $c_{rx} \equiv \bar{X}_r(1-\bar{X}_r)$ and $c_{ry} \equiv \bar{Y}_r(1-\bar{Y}_r)$.

We continue with the batched sampling scheme until the desired half-width criterion is met, with the goal of minimising the number of observations taken along the way subject to the required coverage probability.

What remains is to describe how to select b_r on any given stage. First of all, the CI at the end of stage $r-1$ is given by

$$p_x - p_y \in \bar{X}_{r-1} - \bar{Y}_{r-1} \pm z_{\alpha/2}\sqrt{\frac{c_{r-1,x}}{n_{r-1}} + \frac{c_{r-1,y}}{m_{r-1}}}. \tag{7}$$

In light of the current (stage $r-1$) values of $c_{r-1,x}$ and $c_{r-1,y}$, we will determine the stage-r sample sizes b_r and $B - b_r$ of the X's and Y's, respectively, so as to minimise the expected half-width of the next stage; and because the square root is a monotonic function, this is equivalent to the following surrogate minimisation problem.

$$\begin{aligned} &\text{minimise } \mathrm{E}\left[\frac{c_{rx}}{n_r} + \frac{c_{ry}}{m_r}\right] \text{ subject to } B \\ &\quad \text{stage-}r \text{ observations (and given } c_{r-1,x} \text{ and } c_{r-1,y}). \end{aligned} \tag{8}$$

In order to address this problem, suppose that $\bar{X}_{r-1} = a$, a known quantity at the end of stage $r-1$. Then

$$\begin{aligned} n_r\bar{X}_r = \sum_{i=1}^{n_r} X_i &= \sum_{i=1}^{n_{r-1}} X_i + \sum_{i=n_{r-1}+1}^{n_r} X_i \\ &= n_{r-1}\,a + \sum_{i=n_{r-1}+1}^{n_r} X_i \\ &\sim n_{r-1}\,a + \mathrm{Bin}(b_r, p_x); \end{aligned}$$

and this distributional result implies that

$$
\begin{aligned}
&\mathrm{E}[c_{rx}] \\
&\quad = \mathrm{E}[\bar{X}_r] - \mathrm{E}[\bar{X}_r^2] \;=\; \mathrm{E}[\bar{X}_r] - \mathrm{Var}(\bar{X}_r) - \{\mathrm{E}[\bar{X}_r]\}^2 \\
&\quad = \left[\frac{n_{r-1}a}{n_r} + \frac{b_r p_x}{n_r}\right] - \frac{b_r p_x(1-p_x)}{n_r^2} - \left[\frac{n_{r-1}a}{n_r} + \frac{b_r p_x}{n_r}\right]^2 \\
&\quad \approx \left[\frac{n_{r-1}a}{n_r} + \frac{b_r a}{n_r}\right] - \frac{b_r a(1-a)}{n_r^2} - \left[\frac{n_{r-1}a}{n_r} + \frac{b_r a}{n_r}\right]^2 \\
&\qquad \text{(where we approximate } p_x \text{ by } a\text{)} \\
&\quad = a(1-a)\left[1 - \frac{b_r}{n_r^2}\right] \;=\; c_{r-1,x}\left[1 - \frac{b_r}{n_r^2}\right].
\end{aligned}
$$

Similarly,

$$
\mathrm{E}\left[\frac{c_{rx}}{n_r} + \frac{c_{ry}}{m_r}\right] \approx \frac{c_{r-1,x}}{n_r}\left[1 - \frac{b_r}{n_r^2}\right] + \frac{c_{r-1,y}}{m_r}\left[1 - \frac{B-b_r}{m_r^2}\right].
$$

Under the completely reasonable assumption that the batch size B is small compared to the accumulated actual sample sizes n_r and m_r, we have that $b_r \ll n_r^2$ and $B - b_r \ll m_r^2$; and so we obtain

$$
\mathrm{E}\left[\frac{c_{rx}}{n_r} + \frac{c_{ry}}{m_r}\right] \approx \frac{c_{r-1,x}}{n_r} + \frac{c_{r-1,y}}{m_r}, \tag{9}
$$

which is the intuitively pleasing quantity that we will minimise in light of (8).

To simplify things, let $\gamma_r \in [0,1]$ denote a generic proportion; and define that fraction of the batch size B by the quantity $b_r \equiv \gamma_r B$. Thus, by Equation (9), we have the following problem to solve:

$$
\text{minimise } f(\gamma_r) \equiv \frac{c_{r-1,x}}{n_{r-1} + \gamma_r B} + \frac{c_{r-1,y}}{m_{r-1} + (1-\gamma_r)B}
$$
$$
\text{for } \gamma_r \in [0,1].
$$

We compute the first and second derivatives with respect to γ_r:

$$
f'(\gamma_r) \;=\; \frac{-Bc_{r-1,x}}{(n_{r-1} + \gamma_r B)^2} + \frac{Bc_{r-1,y}}{(m_{r-1} + (1-\gamma_r)B)^2}
$$

$$
f''(\gamma_r) \;=\; \frac{2B^2 c_{r-1,x}}{(n_{r-1} + \gamma_r B)^3} + \frac{2B^2 c_{r-1,y}}{(m_{r-1} + (1-\gamma_r)B)^3}.
$$

We do not consider the case where $c_{r-1,x} = c_{r-1,y} = 0$ since then the half-width is 0. In the case $c_{r-1,x} = 0$, then the minimum value is attained at $\gamma_r = 0$; and when $c_{r-1,y} = 0$, the minimum value is attained at $\gamma_r = 1$. So we consider the case in which $c_{r-1,x}$ and $c_{r-1,y} \neq 0$; and for that we need to solve the equation $f'(\gamma_r) = 0$, which is equivalent to

$$
\begin{aligned}
&Bc_{r-1,x}(m_{r-1} + B(1-\gamma_r))^2 \;=\; Bc_{r-1,y}(n_{r-1} + B\gamma_r)^2 \\
&\quad \Leftrightarrow \sqrt{c_{r-1,x}}(m_{r-1} + B(1-\gamma_r)) \\
&\quad = \sqrt{c_{r-1,y}}(n_{r-1} + B\gamma_r),
\end{aligned}
$$

which follows after a little algebra and the fact that all of the terms are positive. Solving for γ_r we obtain

$$
\gamma_r^\star \;=\; \frac{\sqrt{c_{r-1,x}}(m_{r-1} + B) - \sqrt{c_{r-1,y}}\,n_{r-1}}{B(\sqrt{c_{r-1,x}} + \sqrt{c_{r-1,y}})}.
$$

Notice that depending on the value of B, it may be the case that $\gamma_r^\star$ falls outside of the [0,1] interval. In particular,

- $\gamma_r^\star < 0$ if and only if $B < \sqrt{\frac{c_{r-1,y}}{c_{r-1,x}}}\,n_{r-1} - m_{r-1}$.
- $\gamma_r^\star > 1$ if and only if $B < -n_{r-1} + \sqrt{\frac{c_{r-1,x}}{c_{r-1,y}}}\,m_{r-1}$.

And as B is positive and the right-hand sides of the above inequalities are both 0 or have different signs, then at most one of the two conditions can hold at any time. In fact, if $\gamma_r^\star < 0$, then $f'(\gamma_r) > 0$ for all $\gamma_r \in [0,1]$, and thus the minimum is achieved at $\tilde{\gamma}_r^\star = 0$. Similarly, when $\gamma_r^\star > 1$, then $f'(\gamma_r) < 0$ for all $\gamma_r \in [0,1]$, in which case the minimum is achieved at $\tilde{\gamma}_r^\star = 1$. Finally, for any other case, because $f''(\gamma_r) > 0$ for any $\gamma_r \in (0,1)$, then the location of the minimum is $\tilde{\gamma}_r^\star = \gamma_r^\star$, which occurs in that interval.

These results suggest the following multi-stage procedure using batches: While the desired half-width for the confidence interval has yet to be achieved, compute $\gamma_r^\star$ using the values $c_{r-1,x}, c_{r-1,y}, n_{r-1}, m_{r-1}, B$, and then obtain $\tilde{\gamma}_r^\star$. Compute the number of observations to take in the next stage, b_r and $B - b_r$, by rounding to the nearest integers, $\langle \tilde{\gamma}_r^\star B \rangle$ and $\langle (1 - \tilde{\gamma}_r^\star)B \rangle$, respectively. Take the suggested number of observations from both populations, recalculate the CI (7) with r instead of $r-1$, and stop if the desired half-width is achieved.

This algorithm generalises the individual (one-at-a-time) sampling procedure. By way of motivation, let us consider that $B = 1$ case, where our batch method takes a new observation from the X population if $b_r = \gamma_r^\star B = \gamma_r^\star > 0.5$, i.e., if

$$
\frac{\sqrt{c_{r-1,x}}(m_{r-1} + 1) - \sqrt{c_{r-1,y}}\,n_{r-1}}{\sqrt{c_{r-1,x}} + \sqrt{c_{r-1,y}}} \;>\; 0.5,
$$

which is equivalent to

$$
\frac{c_{r-1,x}}{c_{r-1,y}} \;>\; \frac{n_{r-1}^2 + n_{r-1} + 0.25}{m_{r-1}^2 + m_{r-1} + 0.25}. \tag{10}
$$

Meanwhile, the "original" individual sampling criterion (6) going from stage $r-1$ to stage r is to take an observation from the X population if (after a little algebra)

$$
\frac{c_{r-1,x}}{c_{r-1,y}} \;>\; \frac{n_{r-1}^2 + n_{r-1}}{m_{r-1}^2 + m_{r-1}},
$$

which is roughly the same condition as (10) for moderately large n_{r-1} and m_{r-1}.

5. Monte Carlo analysis

This section presents the results of a Monte Carlo study analysing the performance of the naive two-stage procedure and the sequential procedure with

and without batching. The performance measures under evaluation include the coverage probability, the expected half-width, and the expected number of stages required to obtain the desired half-width. For a large selection of the underlying probabilities p_x and p_y, we consider scenarios involving various choices of the following CI specifications: maximum desired half-width ϵ; initial (equal) sample sizes $n_1 = m_1$; and batch size B. All scenarios are simulated with 1000 replications, with common random numbers used for the observations, to allow for "apples-to-apples" comparisons among procedures. In order to obtain nontrivial (nonzero) initial estimators $\sqrt{c_{1x}}$ and $\sqrt{c_{1y}}$ of the sample standard deviations of the X's and Y's, respectively, our initial sample sizes occasionally had to be increased to some multiple of the "tentative" initial n_1 – to ensure at least one success and failure from the X's and Y's; this contingency was more likely to be necessitated when p_x or p_y were very close to 0 or 1. Notice that because of the common random numbers, the initial sample for a particular simulation replication was always the same for all procedures, and otherwise the comparisons between procedures were not affected.

5.1. Two-stage procedure results

Figure 3 presents estimates of the coverage and expected number of observations for the naive two-stage procedure with tentative initial sample sizes of $n_1 = m_1 = 35$; desired half-widths $\epsilon = 0.01$ and 0.02; and confidence level $1 - \alpha = 0.95$. In order to compute the results, representative choices of the probability pairs ("scenarios") (p_x, p_y) were selected and presented along the horizontal axis. Subfigure (a) at the top presents the following probability pairs from left to right ($p_x \leq 0.35$):

- $p_x = 0.01,\ p_y = 0.01$
- $p_x = 0.03,\ p_y = 0.01, 0.03$
- $p_x = 0.05,\ p_y = 0.01, 0.03, 0.05$
- $p_x = 0.1,\ p_y = 0.01, 0.03, 0.05, 0.1$
- $p_x = 0.2,\ p_y = 0.01, 0.03, 0.05, 0.1, 0.2$
- $p_x = 0.35,\ p_y = 0.01, 0.03, 0.05, 0.1, 0.2, 0.35$

The bottom subfigure (b) presents the following probability pairs from left to right ($p_x \geq 0.5$):

- $p_x = 0.5,\ p_y = 0.01, 0.03, 0.05, 0.1, 0.2, 0.35, 0.5$
- $p_x = 0.75,\ p_y = 0.01, 0.03, 0.05, 0.1, 0.2, 0.35, 0.5, 0.75$
- $p_x = 0.9,\ p_y = 0.01, 0.03, 0.05, 0.1, 0.2, 0.35, 0.5, 0.75, 0.9$

Figure 4 illustrates the same results, but for tentative initial sample sizes of $n_1 = m_1 = 100$.

We are first interested in how well the two-stage method covers the parameter of interest, $p_x - p_y$. Thus, for a particular subfigure, the red and orange lines give the values of the coverages achieved by the two-stage method for $\epsilon = 0.01, 0.02$, respectively. Second, we are concerned with the expected total number of observations (from either X or Y) resulting from our request to obtain a desired CI half-width ϵ; this is given by the blue and purple plots ($\epsilon = 0.01, 0.02$, respectively) in a particular subfigure.

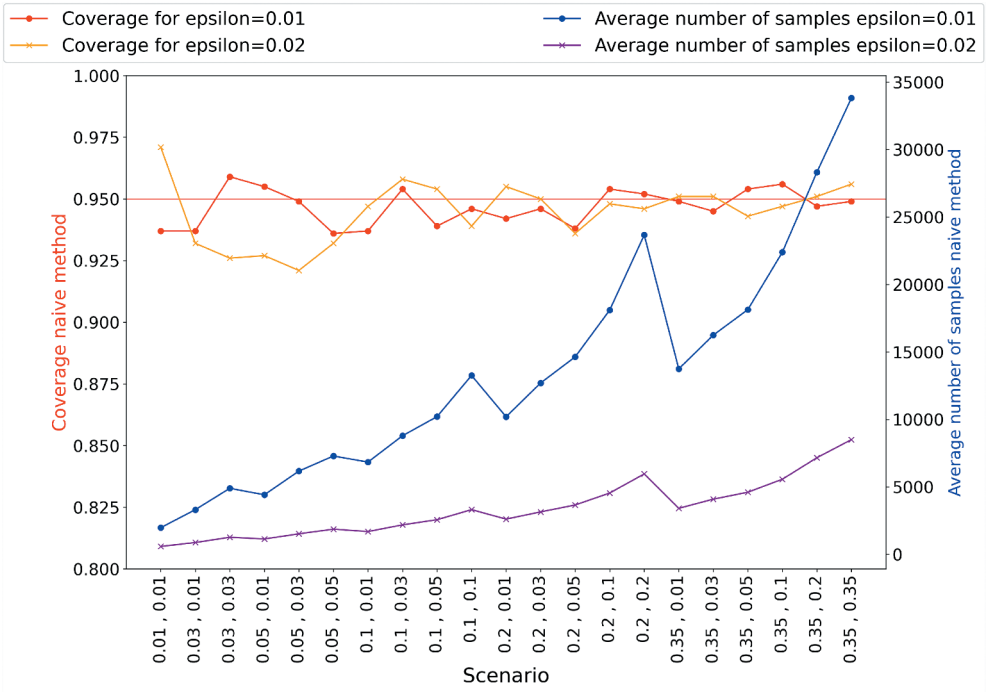
For a given fixed p_x, it is easy to see that the number of observations required by the two-stage procedure increases when p_y increases, and the maximum value is attained when both parameters are equal. Now, when we let p_x increase from 0.01, 0.03, ..., 0.5, we observe an increasing cyclic pattern up to $p_x = 0.5$ that has local peaks at the scenarios for which $p_x = p_y$; and then a decreasing cyclic pattern, this time with maximum local values when $p_y = 0.5$. These two qualitative phenomena occur for the various choices of the initial sample sizes and the desired half-widths under study in the current paper.

It is often the case that the coverage of 95% is not achieved, though the coverage usually seems to be approximately the same for both choices of ϵ. In fact, the coverage actually falls well below 95% when the initial sample is small ($n_1 = m_1 \leq 35$) and at least one of p_x or p_y is large (≥ 0.9). (Note that the coverage estimates generally have standard errors of about $\sqrt{(0.95)(0.05)/1000} = 0.0069$.) We also remark that the average number of observations required for fixed ϵ is not particularly affected by the choice of the initial sample size, at least in our examples, where $n_1 = m_1$ is small compared to the final average number of observations required.

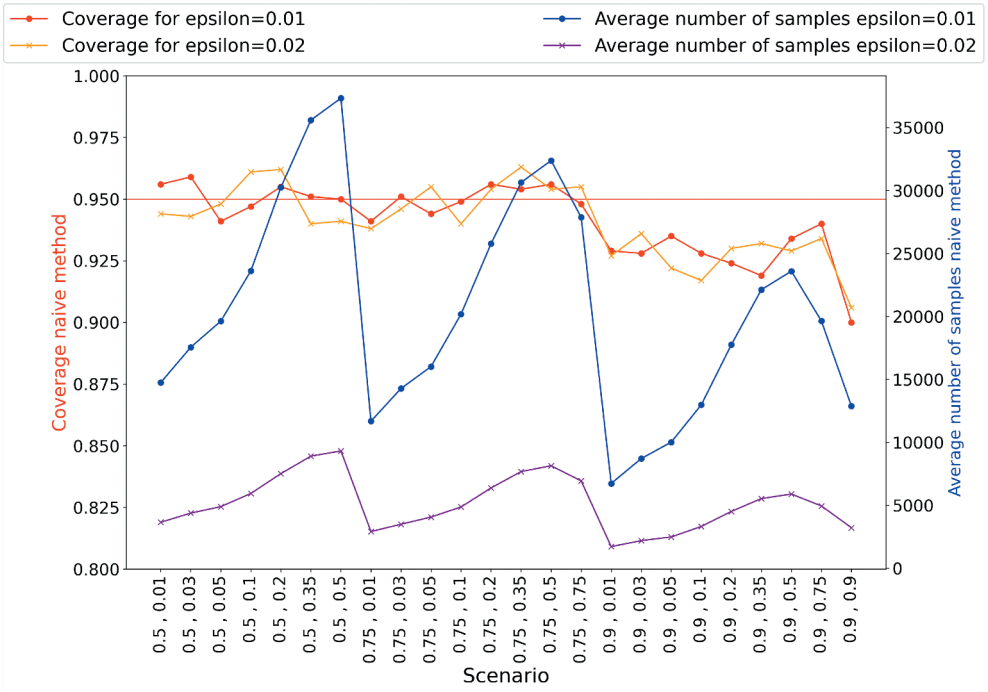
5.2. Two-stage procedure vs. sequential procedure

We now use Monte Carlo simulation to compare the performance of the naive two-stage and one-at-a-time (fully) sequential procedures. Our work proceeds in parallel to our discussion in §5.1, except we are now concerned with *differences* in the performance between the two procedures.

The Monte Carlo results are again based on a tentative initial sample of n_1 observations from both populations. In particular, Figures 5–7 correspond to $n_1 = 10$, 35, and 100, respectively: all with $1 - \alpha = 0.95$. Each figure is itself comprised of two subfigures, corresponding to desired half-widths $\epsilon = 0.01, 0.02$. Each subfigure depicts the same sets of results for representative choices of the probability pairs (scenarios) (p_x, p_y) along the horizontal axis (the same choices as were used in §5.1). For a specific subfigure, the red and orange plots give values of the *differences* in coverages achieved by the two-stage and the one-at-a-time



(a) Probability pairs, $p_x \leq 0.35$

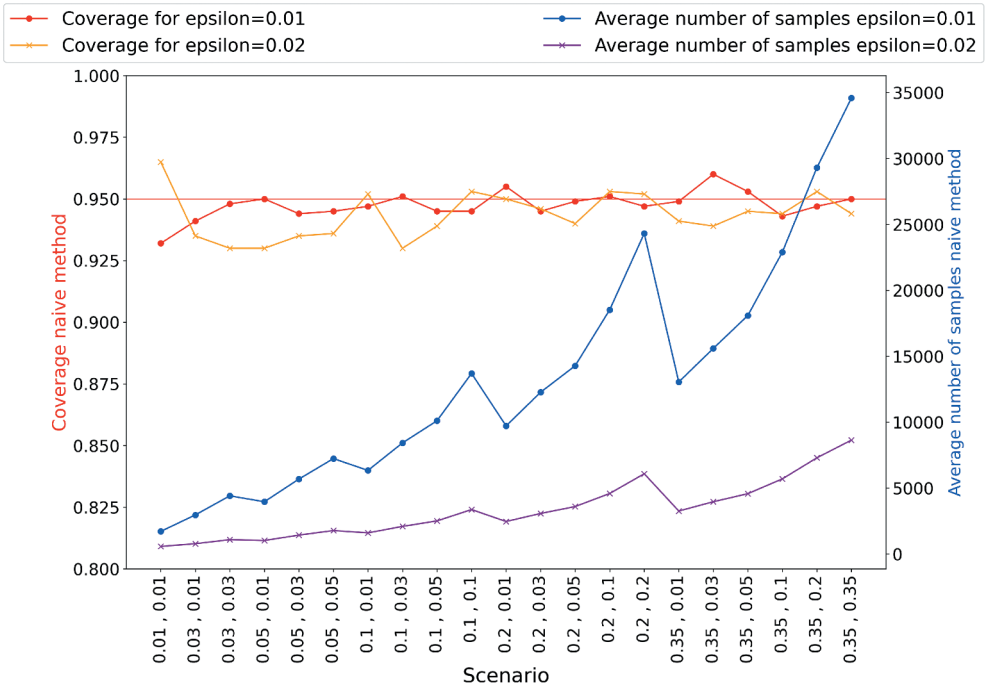


(b) Probability pairs, $p_x \geq 0.5$

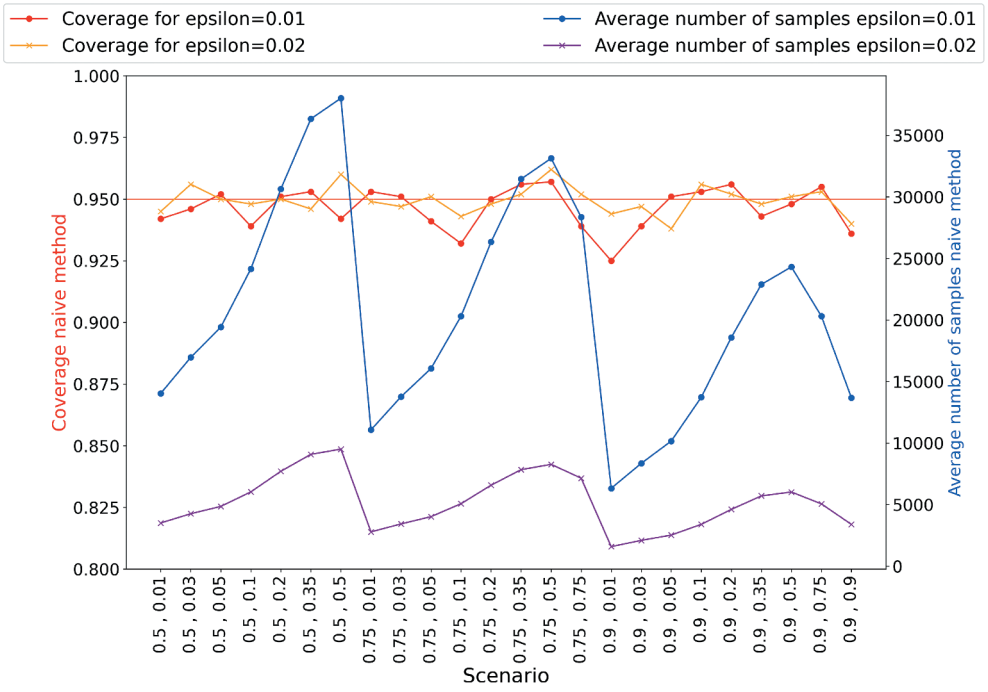
Figure 3. Naive two-stage procedure: Estimated coverage and expected number of observations for various probability pairs (p_x, p_y); $n_1 = m_1 = 35$; $\epsilon = 0.01, 0.02$; and $1 - \alpha = 0.95$.

sequential procedures. The blue and purple plots in a particular subfigure compare the *differences* in expected sample sizes between the naive two-stage and the one-at-a-time sequential procedures.

Figure 5 gives results for which the tentative initial sample is of size $n_1 = 10$. For each choice of parameters (p_x, p_y) the coverage difference seems to be small and not much affected even by very different values of p_x and p_y. (It is not possible to reject the hypothesis that the coverage differences are statistically insignificant as the coverage difference estimates generally have standard errors of about $\sqrt{2(0.95)(0.05)/1000} = 0.0097$.) On the other hand, we observe significant cyclic behaviour for the sample-size results, indicating scenarios in which one method is more parsimonious than the other. When one of p_x



(a) Probability pairs, $p_x \leq 0.35$

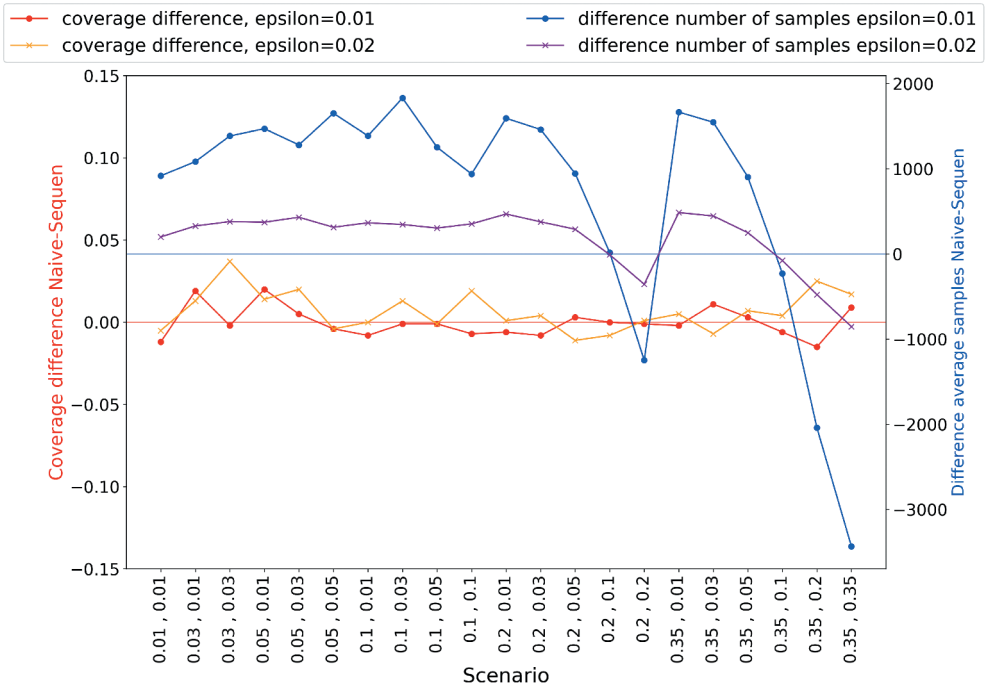


(b) Probability pairs, $p_x \geq 0.5$

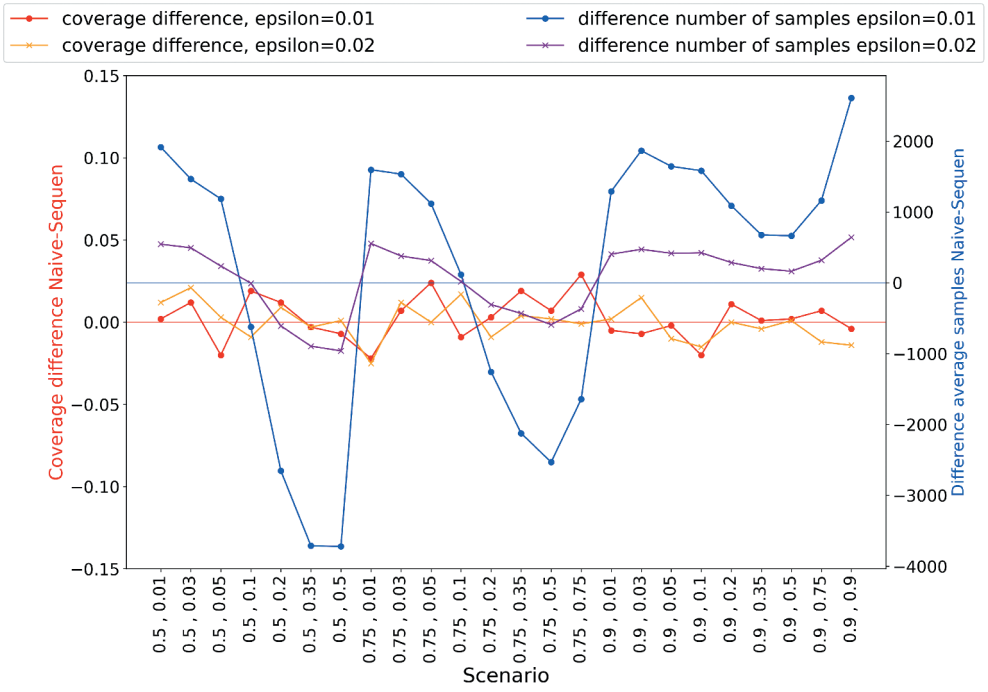
Figure 4. Naive two-stage procedure: Estimated coverage and expected number of observations for various probability pairs (p_x, p_y); $n_1 = m_1 = 100$; $\epsilon = 0.01, 0.02$; and $1 - \alpha = 0.95$.

and p_y is at most 0.05, the naive two-stage method performs worse than the fully sequential; and when $p_x = 0.9$ the two-stage procedure also requires more samples. For all other scenarios the naive two-stage procedure performs at least as well as the fully sequential procedure with respect to sample size; and this advantage is greater when both probabilities are close to each other, likely due to difficulty in the sequential procedure's ability to distinguish between two probabilities that are approximately equal. Of course, the underlying reason for the two-stage procedure's good sample-size performance is that it cannot guarantee the desired half-width – a major drawback of that method.

Figure 6 gives analogous results for tentative initial sample size $n_1 = 35$. Now we find a smaller



(a) Probability pairs, $p_x \leq 0.35$

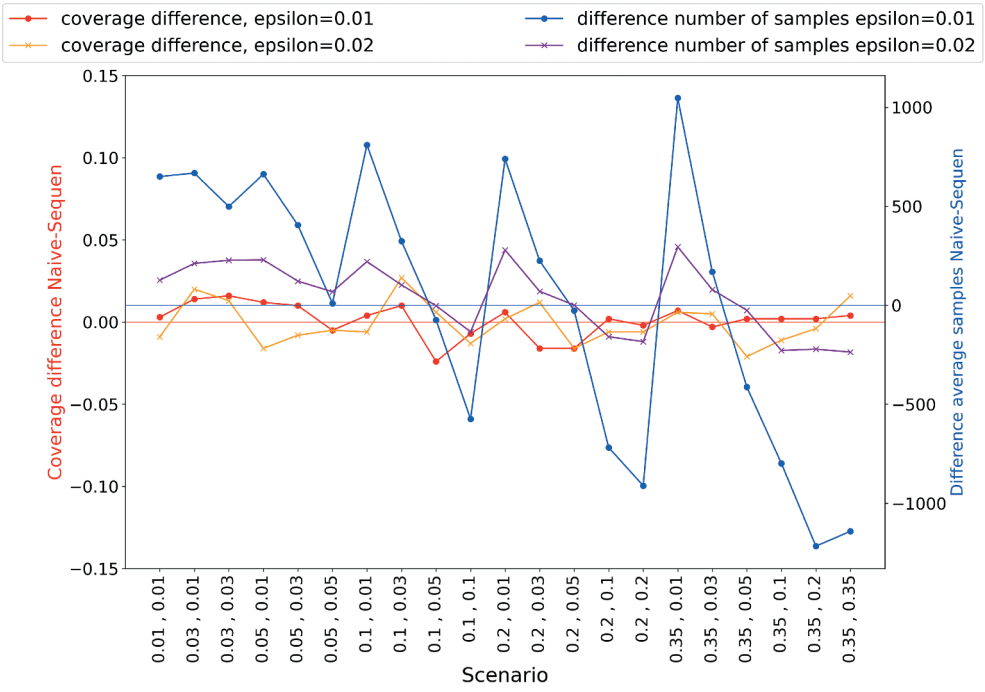


(b) Probability pairs, $p_x \geq 0.5$

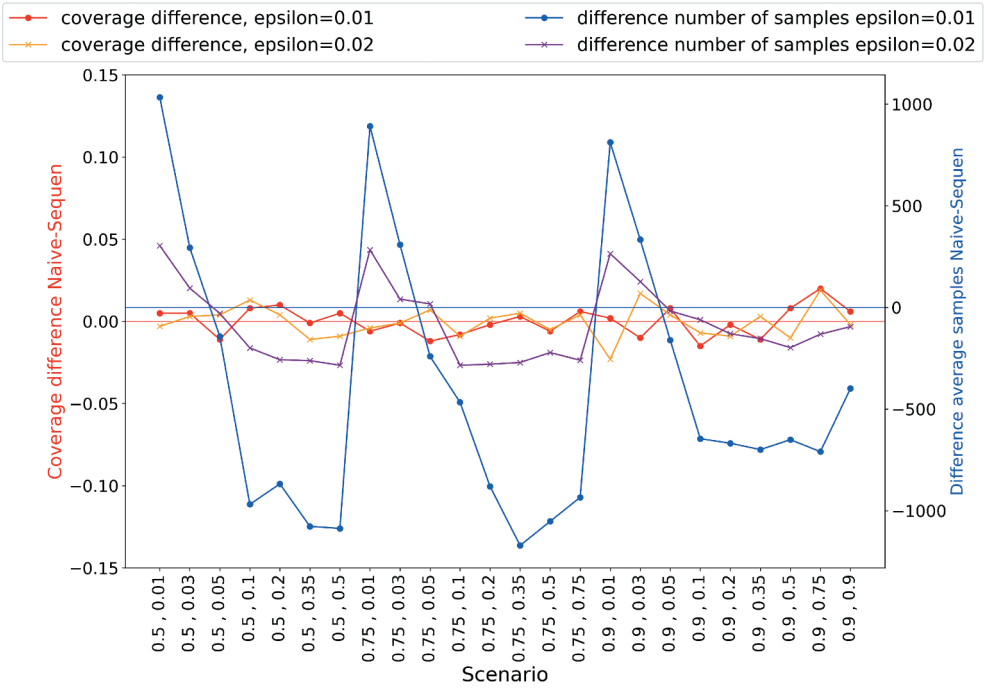
Figure 5. Naive two-stage procedure vs. fully sequential procedure: Coverage and sample-size differences for initial sample sizes $n_1 = m_1 = 10$; half-widths $\epsilon = 0.01, 0.02$; and $1 - \alpha = 0.95$.

set of probability pairs (p_x, p_y) for which the fully sequential method outperforms the two-stage procedure in terms of sample size. Here the sample-size advantage of the fully sequential method is restricted to probability pairs where both p_x, p_y are at most 0.05 and to pairs where one of p_x, p_y is very small (at most 0.03). For all other scenarios, the two-stage procedure performs better in terms of sample size, and its advantage is greater when $p_x \approx p_y$. We also see that the performance differences between the methods are quantitatively smaller when we use an initial sample size of $n_1 = 35$ compared to $n_1 = 10$ (everything else being held fixed).

The trend just described continues in Figure 7. Now the fully sequential method has a sample-size



(a) Probability pairs, $p_x \leq 0.35$



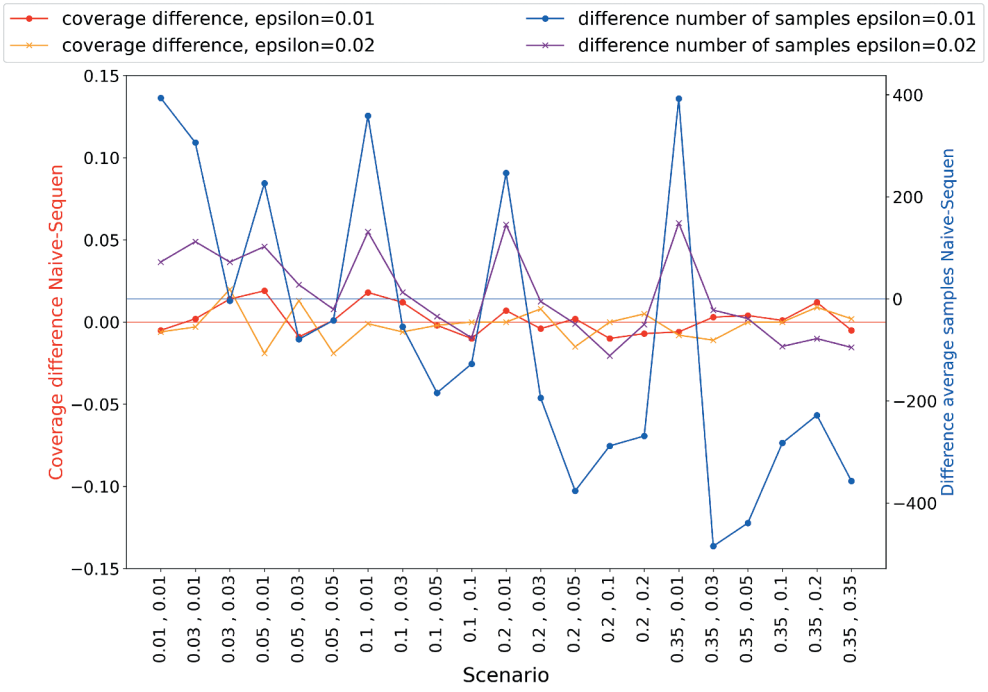
(b) Probability pairs, $p_x \geq 0.5$

Figure 6. Two-stage procedure vs. fully sequential procedure: Coverage and sample-size differences for initial sample sizes $n_1 = m_1 = 35$; half-widths $\epsilon = 0.01, 0.02$; and $1 - \alpha = 0.95$.

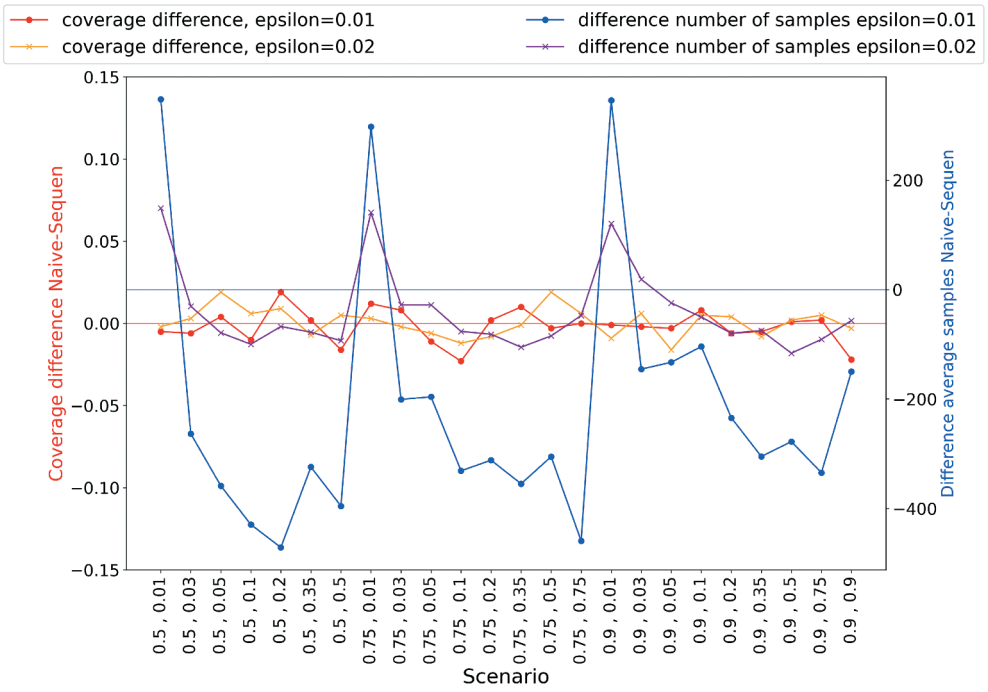
advantage over the two-stage method only when at least one of p_x, p_y is equal to 0.01 (at least for the scenarios considered herein). Moreover, the performance differences between the methods are smaller compared to the $n_1 = 10$ and 35 cases. Yet again, the sample-size victory for the two-stage method rings a bit hollow since that procedure does not guarantee our half-width requirement.

5.3. Fully sequential vs. batches

We evaluate the performance of the sequential method using batches of size $B \geq 1$. Batching will clearly result in



(a) Probability pairs, $p_x \leq 0.35$

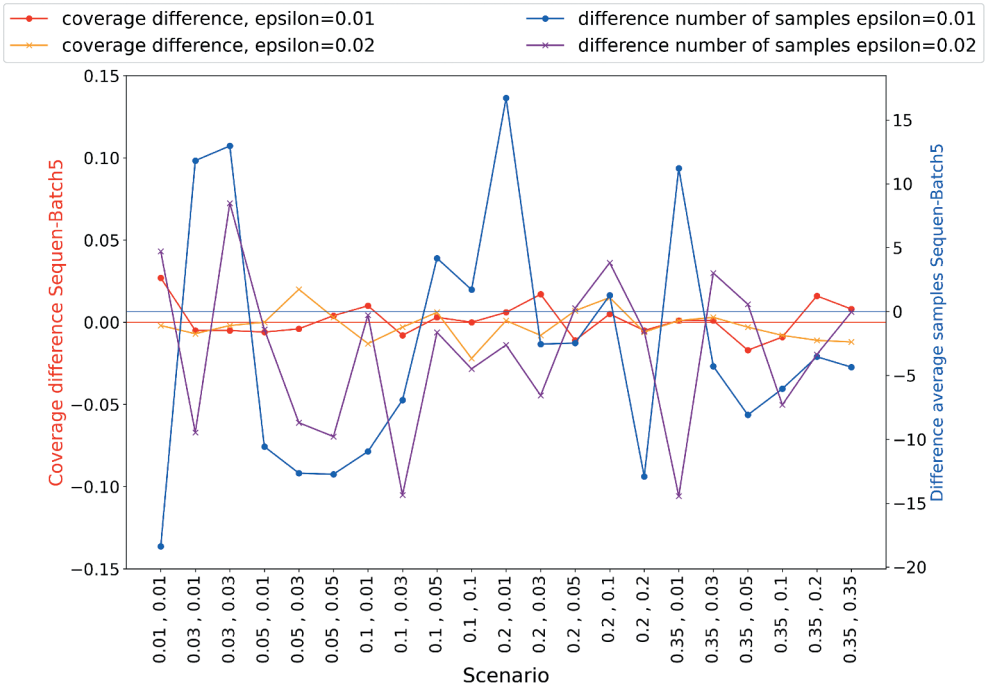


(b) Probability pairs, $p_x \geq 0.5$

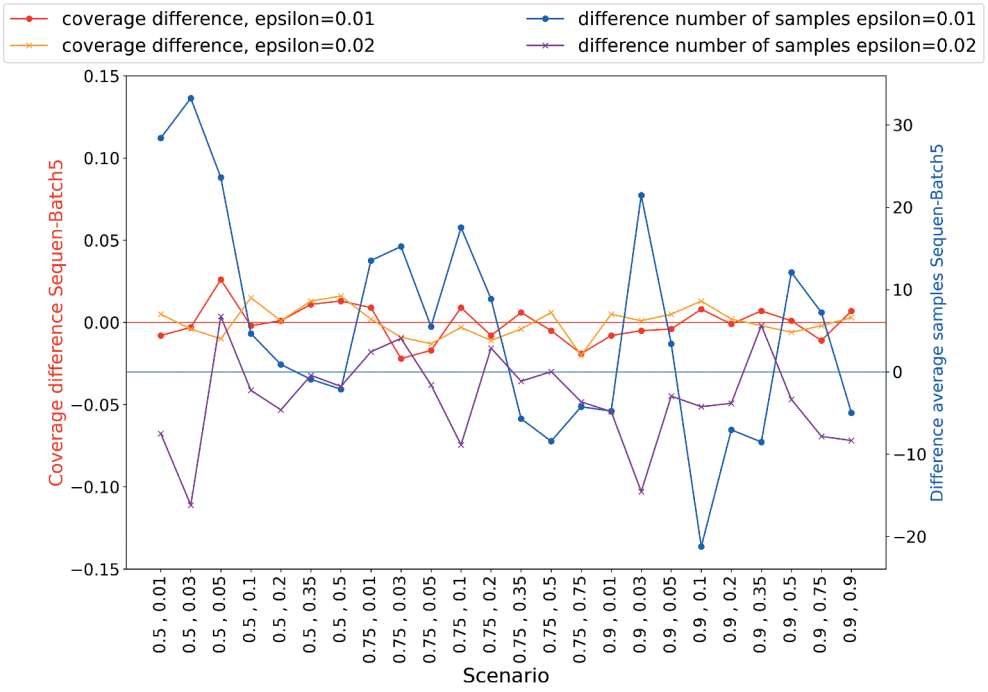
Figure 7. Two-stage procedure vs. fully sequential procedure: Coverage and sample-size differences for initial sample sizes $n_1 = m_1 = 100$; half-widths $\epsilon = 0.01, 0.02$; and $1 - \alpha = 0.95$.

a smaller number of *stages* compared to the one-at -a-time ($B = 1$) procedure – since stages roughly correspond to time expenditures. Thus, we will try to select the "best" batch size $B^\star$ to balance the time savings achieved by batching against any deterioration in coverage or increase in total sample size. This is perhaps a win-win situation, because for small batch sizes B, we intuitively expect coverage and total sample-size results to be roughly the same as those arising from the $B = 1$ case.

For ease of exposition, we discuss results for batches of sizes $B = 5$, 10, and 20. In order to compare the batched procedure vs. the fully sequential procedure, we simulated the previous 45 (p_x, p_y) scenarios, using common random numbers as before. As expected, when B increased, the total number of observations required increased on average, but just modestly. It turns out that moving from batch sizes of 5 to 10 to 20 did not produce a clear change trend with



(a) Probability pairs, $p_x \leq 0.35$

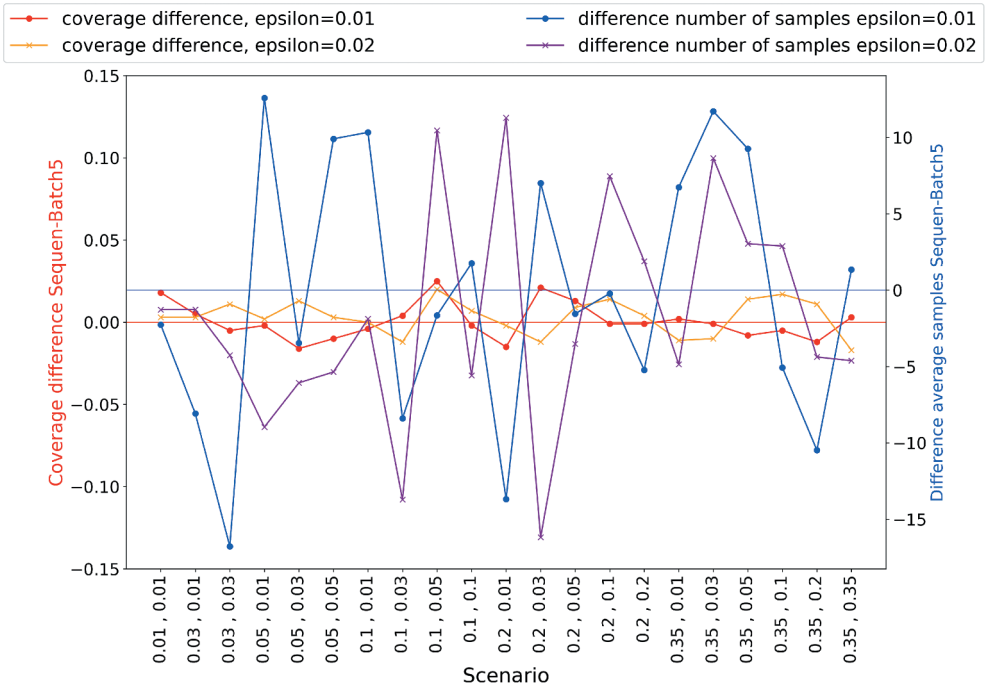


(b) Probability pairs, $p_x \geq 0.5$

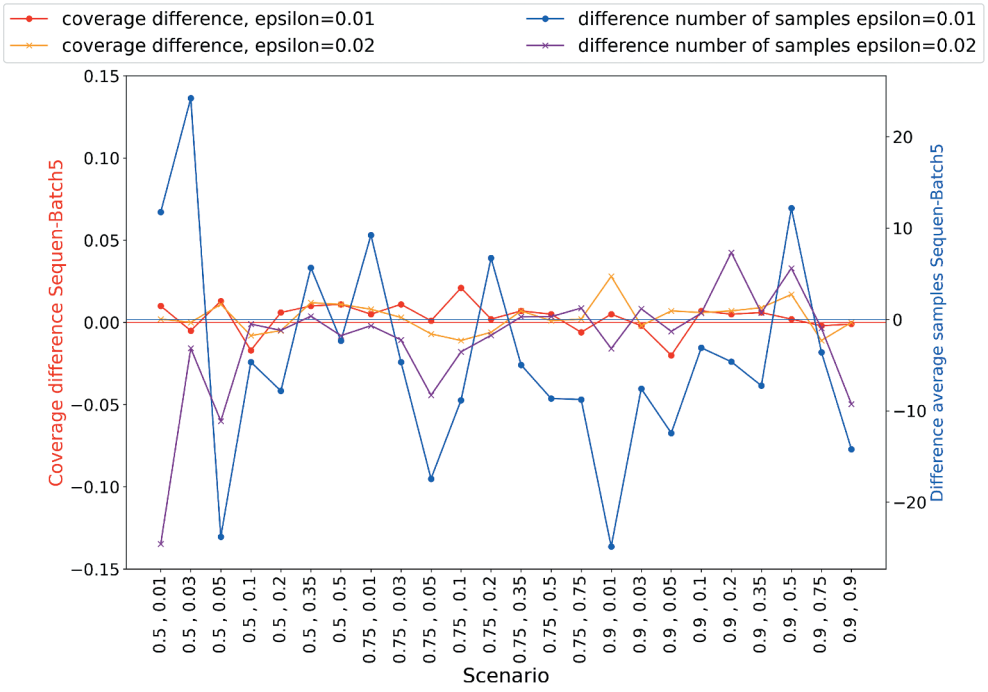
Figure 8. One-at-a-time fully sequential procedure vs. batched procedure: Coverage and sample-size differences for initial sample sizes $n_1 = m_1 = 10$; batch size $B = 5$; half-widths $\epsilon = 0.01, 0.02$; and $1 - \alpha = 0.95$.

respect to the coverage probability; and so we shall only give here a representative comparison for the case $B = 5$.

Figure 8 illustrates the difference in performance between the one-at-a-time ($B = 1$) fully sequential procedure and the batch procedure with $B = 5$, both incorporating an initial sample of tentative size 10. For most of the 45 scenarios, the numbers of observations required by the batch method are approximately equal to those required by the one-at -a-time sequential method (the figure merely shows noise around the horizontal blue line at 0); and the



(a) Probability pairs, $p_x \leq 0.35$



(b) Probability pairs, $p_x \geq 0.5$

Figure 9. One-at-a-time fully sequential procedure vs. batched procedure: Coverage and sample-size differences for initial sample sizes $n_1 = m_1 = 35$; batch size $B = 5$; half-widths $\epsilon = 0.01, 0.02$; and $1 - \alpha = 0.95$.

coverages are also approximately equal for the $B = 1$ and $B = 5$ cases. We also note that the differences between the $B = 1$ and $B = 5$ cases in terms of total observations are typically small (compared to the differences between the fully sequential $B = 1$ method vs. the naive two-stage procedure, as discussed earlier). Figure 9 shows the same comparison, but for initial samples of tentative size $n_1 = 35$. Under this scenario it is easy to see that the coverage is also preserved, but now the batching method seems to need just a few extra observations compared to the fully sequential procedure (as more dots lie below the

horizontal blue line at 0). This very minor increase in the number of observations is almost certainly worth the trouble if "time" is considered as a performance measure – because $B = 5$ requires only about 1/5 of the stages needed by the fully sequential $B = 1$ method.

6. Conclusions

This paper analysed various versions of the classical confidence interval for the difference between two Bernoulli success parameters, namely, fixed-sample-size, two-stage, fully sequential, and batched-sequential procedures. The analysis was undertaken via exact and Monte Carlo methods, where we examined the attained coverage, attained half-width, and required sample sizes. Our procedures generally conducted sampling in an intelligent way, in that we proposed a greedy heuristic (based on the estimated sample standard deviations of the two populations) that aimed to reduce the CI half-width as much as possible from stage to stage.

We found that for very small sample sizes, the various procedures under study often missed their target coverage, either undershooting or overshooting – this is a well-known phenomenon arising from the discreteness of the two Bernoulli populations. As the sample sizes increased, the coverages were often slightly below nominal for the sequential procedures (also a well-known phenomenon); but as the sample sizes increased further, then the nominal coverages were, for all practical purposes, obtained. Our intelligent heuristic procedures achieved (statistically) the same coverage as the corresponding optimised two-stage procedure, yet with smaller expected sample sizes for some of the probability pairs (p_x, p_y), particularly those heavily unbalanced; and the batched-sequential procedure saves on the anticipated number of stages at the cost of only a relatively insignificant increase in the total number of observations.

As we mentioned above, it is often the case that for small sample sizes, Bernoulli CIs undershoot or overshoot the intended coverage. Future research will consider elementary corrections along the lines of Agresti and Coull (1998) to mitigate these issues. In addition, one can also improve small-sample performance if one has reasonable (i.e., at least "rough") information about the values of p_x and p_y before sampling begins; to this end, one could adopt a Bayesian approach or the "tandem" approach taken by Yaacoub et al. (2019a) for the one-parameter CI case. In addition, we will study true optimal procedures in the spirit of the one-dimensional methodology of Yaacoub et al. (2019b). Finally, in many applications, particularly in the context of discrete-event computer simulation, one is interested in the case in which the components of the pair (X_i, Y_i) are correlated. For instance, determine which of two competing inventory policies is more likely to fulfill an order on-time – a case that can be tested via simulation using the same customers thanks to common random numbers. This is the subject of an ongoing sister paper.

Acknowledgments

The authors thank the Associate Editor and two referees for their thoughtful comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician*, *52*, 119–126.

Armitage, P. (1958). Numerical studies in the sequential estimation of a binomial parameter. *Biometrika*, *45*(1–2), 1–15. https://doi.org/10.1093/biomet/45.1–2.1

Bechhofer, R. E., Santner, T. J., & Goldsman, D. (1995). *Design and analysis of experiments for statistical selection, screening and multiple comparisons*. New York: John Wiley and Sons.

Brown, L. D., Cai, T., & Dasgupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, *16*(2), 101–133. https://doi.org/10.1214/ss/1009213286

Brown, L. D., Cai, T., & Dasgupta, A. (2002). Interval estimation for a binomial proportion and asymptotic expansions. *Annals of Statistics*, *30*(1), 160–201. https://doi.org/10.1214/aos/1015362189

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*(4), 404–413. https://doi.org/10.1093/biomet/26.4.404

Frey, J. (2010). Fixed-width sequential confidence intervals for a proportion. *The American Statistician*, *64*(3), 242–249. https://doi.org/10.1198/tast.2010.09140

Goldsman, D. (2015). A practical guide to ranking and selection methods. In D. M. Aleman & A. C. Thiele Eds., *The Operations Research revolution* (pp. 89–110). Institute for Operations Research and the Management Sciences. *TutORials in Operations Research* series, ed. J. C. Smith. Cantonville, MD: INFORMS.

Huber, M. (2017). A Bernoulli mean estimate with known relative error distribution. *Random Structures & Algorithms*, *50*(2), 173–182. https://doi.org/10.1002/rsa.20654

Khan, R. A. (1998). Fixed-width confidence sequences for the normal mean and the binomial probability. *Sequential Analysis*, *17*(3–4), 205–217. https://doi.org/10.1080/07474949808836409

Malinovsky, Y., & Zacks, S. (2018). Proportional closeness estimation of probability of contamination under group

testing. *Sequential Analysis*, *37*(2), 145–157. https://doi.org/10.1080/07474946.2018.1466518

Mukhopadhyay, N., & Banerjee, S. (2015). Purely sequential and two-stage bounded-length confidence intervals for the Bernoulli parameter with illustrations from health studies and ecology. In P. Choudhary, C. Nagaraja, & H. K. T. Ng (Eds.), *Ordered data analysis, modeling and health research methods. In Honor of H. N. Nagaraja's 60th Birthday* (pp. 211–234). New York: Springer.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, *17*(8), 857–872. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E"3.0.CO;2-E"https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E

Robbins, H., & Siegmund, D. (1974). Sequential estimation of *p* in Bernoulli trials. In E. J. G. Pitman & E. J. Williams (Eds.), *Studies in probability and statistics* (pp. 103–107). Jerusalem Academic Press.

Tanaka, M. (1961). On a confidence interval of given length for the parameter of the binomial and the Poisson distributions. *Annals of the Institute of Statistical Mathematics*, *13*(1), 201–215. https://doi.org/10.1007/BF02868870

Turner, A. J., Balestrini-Robinson, S., & Mavris, D. (2013). Heuristics for the regression of stochastic simulations. *Journal of Simulation*, *7*(4), 229–239. https://doi.org/10.1057/jos.2013.1

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212. https://doi.org/10.1080/01621459.1927.10502953

Yaacoub, T., Goldsman, D., Mei, Y., & Moustakides, G. V. (2019a). Tandem-width sequential confidence intervals for a Bernoulli proportion. *Sequential Analysis*, *38*(2), 163–183. https://doi.org/10.1080/07474946.2019.1611315

Yaacoub, T., Moustakides, G. V., & Mei, Y. (2019b). Optimal stopping for interval estimation in Bernoulli trials. *IEEE Transactions on Information Theory 65* (5), 3022–3033. https://doi.org/10.1109/TIT.2018.2885405

Zacks, S., & Mukhopadhyay, N. (2007). Distributions of sequential and two-stage stopping times for fixed-width confidence intervals in Bernoulli trials: Application in reliability. *Sequential Analysis*, *26*(4), 425–441. https://doi.org/10.1080/07474940701620907